An upgrade and revision of the chimpanzee reference genome

Lukas Kuderna [1], Chad Tomlinson [2], Andrew J Sharp [3], Lars Feuk [4], Richard E. Green [5], Wesley C Warren [2], Tomas Marques-Bonet [1,6]

[1] Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain, [2] The Genome Institute, Washington University School of Medicine, St. Louis, USA, [3] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA, [4] Department of Immunology, Genetics and Pathology, Rudbeck Laboratory and Science for Life Laboratory, Uppsala University, Uppsala, Sweden, [5] Department of BiomolecularEngineering, University of California, Santa Cruz, CA 95060, USA., [6] Centro Nacional de Análisis Genómico (Parc Cientific de Barcelona), Baldiri Reixac 4, 08028 Barcelona, Spain.

Introduction

Comparative and evolutionary genomics heavily rely on reference genome assemblies, howeverall but two mammalian references - human and mouse - are based on a whole genome shotgun (WGS) assembly and thus considered draft representations. This also applies to the chimpanzee (*Pan troglodytes*), a species with a heavily fragmented reference assembly yet of key importance to study our own evolutionary trajectories [1]. Here, we applied different novel sequencing and assembly strategies to improve the current status of the chimpanzee genome. Our the goal to improve the reference assembly and reassess the chimpanzee's position in the context of great ape evolution.

Methods

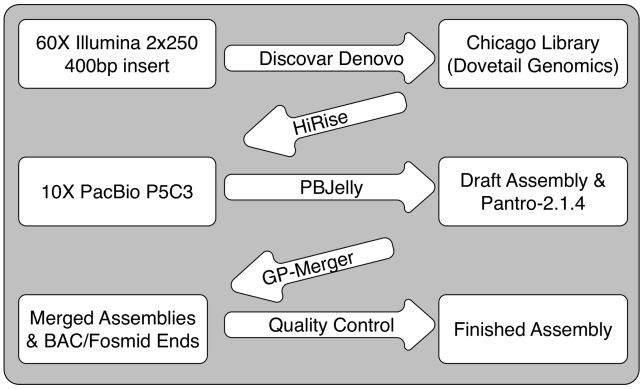

Figure 1: A schematic overview of our assembly pipeline

We assembled 60X of Illumina 2x250 paired end reads with Discovar Denovo [2]. Contigs from this assembly were scaffolded with a Chicago (*cell free Hi-C for assembly and genome organisation*) library [3], a method that takes advantege of mid range chromosomal interactions to sticht together contigs. Gaps in the scaffolded assembly were filled [4] with long reads sequenced on the Pacific Biosciences RSI platform. Regions from Pantro-2.1.4 (the current chimpanzee reference) spanning remaining gaps were merged in. BAC and Fosmid libraries were used to estimate sizes of remaining gaps and asses misassembly events. All data is derived from the same cell line.

Results

Our new assembly constitutes a significant improvement over the current one, increasing continuity by over 500% atthe contig level , and by almost 300% at the scaffold level (defined via the N50 statistic, see Table 1).

	Pantro-2.1.4	Pantro5V0.2 preliminary
Number of scaffolds	24,129	45,000
Scaffold N50 (bp)	8,925,874	26,673,241
Number of contigs	148,553	72,817
Contig N50 (bp)	64,231	334,510

Table 1: Comparison of assembly metrics of the current vs. our new assembly.

We performed whole genome alignments to Pantro-2.1.4 and found our assembly to align over 2.74Gbp (95%) with an identity of 99.8% and thus well within a range that can be expected due to heterozygosity. We furthermore checked for misassembly events with existing BAC-end libraries (see Table 2). We find 87% (n=45922) of BACs having both ends mapped onto the same scaffold, and of those, almost 99% in expected orientation with only 0.2% having an insert size out of the expected range. These results further validate the quality of our assembly (see Figure 2). We furthermore reduce the total number of contigs, whose fragmentation is one of the main sources for missing or erronous gene models [5], by over a half from 148,553 to 72,817.

Orientation of mapped end	Count (Percentage)
+/+	221 (0.44%)
+/-	45922 (98.77%)
-/+	120 (0.26%)
-/-	242 (0.52%)

Table 2: BAC-end mappings as quality control show that the majority of our assembly is free of large misassemblies

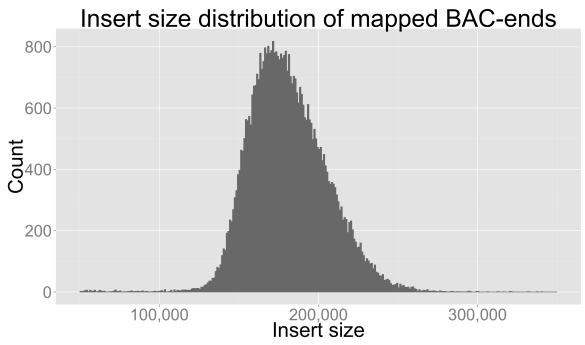

Figure 2: Insert size distribution of mapped BAC-ends on the new reference shows that most BACs map wioth an insert size in the expected range

Conclusions

We deploy a very cost effective approach to construct a highly improved reference genome for the chimpanzee. We therefore do not only validate these novel strategies that may prove useful for *de novo* assemblies of other species, but mainly also offer a new opportunity for a second genomic look at our closest living relative.

Bibliography

1. Mikkelsen et al. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. doi:10.1038/nature04072
2. Available at http://www.broadinstitute.org/software/discovar/blog/
3. Putnam et al. (2015). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *bioRxiv*, 1–25
4. English et al (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One*, 7(11), e47768. doi:10.1371/journal.pone.0047768
5. Denton, J. F. *et al*. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLoS Comput. Biol.* **10,** e1003998 (2014).